Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Testing the applicability of machine learning strength models for stabilized dredged sludge and slurry to new materials

Lu Shen
Jiangxi University of Science and Technology
Tugen Feng
tgfeng75@163.com
Jiangxi University of Science and Technology
Ming Ma
Wanjiang University of Technology

Article

Keywords: dredged sludge, unconfined compressive strength, machine learning, grouped cross-validation, model calibration, water content

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11469847/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Testing the applicability of machine learning strength models for stabilized dredged sludge and slurry to new materials

Lu Shen[1,2,3], Tugen Feng[1,3,*], Ming Ma[1,2]

[1] School of Civil Engineering, Jiangxi University of Science and Technology, Ganzhou 341000, Jiangxi Province, China.

[2] Wanjiang University of Technology, Ma'anshan 243031, Anhui Province, China.

[3] Key Laboratory for Geotechnical Engineering of Ministry of Water Resources, Hohai University, Nanjing 210098, Jiangsu Province, China.

* Corresponding author: Tugen Feng, Professor, Ph.D. E-mail: tgfeng75@163.com

Abstract

Literature-trained machine learning models support mixture design for stabilized dredged sludge and slurry, but random record splits test them on studies already in training rather than on new materials. We combined leave-one-source-out (LOSO) validation, a source-coverage comparison, within-material water-content pairs and comparisons of calibrated and local models. Models trained on 399 unconfined compressive strength records from 10 sources were tested on four new materials and replicated with alkali-activated concrete (AAC) data. The random forest's LOSO mean absolute error (MAE) was 3.74 times its record-random-split median. Replacing the predicted source's training records with an equal number from other sources raised the error 2.57-fold (1.77-fold for AAC). A random forest using stabilizer dosage, age and water content had lower LOSO error than one using all inputs but predicted zero or less accurate within-material strength changes for some materials. With 10 local records, the calibrated random forest had a lower median MAE than raw-scale local ridge regression for two of three materials, whereas Gaussian process regression, best under LOSO validation, did so for none. For a new material, error should be assessed by LOSO validation, water-content effects by within-material pairs and calibration against local models on the same records.

Keywords: dredged sludge; unconfined compressive strength; machine learning; grouped cross-validation; model calibration; water content

1 Introduction

Dredged sludge and engineering waste slurry are fine-grained wastes produced in large volumes with high water contents [1–4]. After stabilization with cement, lime or fly ash (hereafter stabilized fine-grained materials), they are reused as subgrade fill and flowable fill [4–7]. Because the properties of the raw material vary with its origin, the mixture for each new material is set by laboratory strength tests at several dosages and curing ages [2,5,6]. Unconfined compressive strength (UCS) is the main design criterion for the mixture proportions of these materials [2,5,6] and depends on the water content of the raw material, the type and dosage of the stabilizer and the curing conditions [1,8]. Machine learning models predict UCS from variables such as stabilizer dosage and soil properties [9–12] and are used to support mixture design and reduce laboratory testing. Kumar et al. [10] and Aggarwal et al. [11] trained their models on data from a single study. Onyelowe et al. [9] and M'harzi Alaoui et al. [12] instead trained models on data compiled from the literature (hereafter source models), which widened the range of materials represented in training. When a source model is applied to a new material, none of that material's records are in the training data, whereas the reported accuracy of source models was obtained with test records from studies also used in training [12].

The test error of a source model should be obtained under the conditions of its intended use, namely prediction for a material that is absent from the training data. M'harzi Alaoui et al. [12] split their records at random into training (80%) and test (20%) sets. Records from one study share raw materials and specimen preparation methods, and the same mixture is tested at several curing ages. Because a record-based split can place records of the same mixture and the same study in both the training and test sets, its test accuracy reflects prediction for sources already represented in training. A mixture-group split keeps all records of a mixture on the same side of the split. Leave-one-source-out (LOSO) validation holds out all records of one source at a time (Section 2.1 describes how records were grouped into sources), which reproduces the conditions of predicting a new material [13–15]. Prediction errors increase when models are applied to materials or studies outside

the training set [16,17]. Under leave-one-study-out validation of geopolymer strength models, R^2 exceeded 0.98 for studies with at least 10 samples, whereas it fell to large negative values for studies with four or fewer [18]. For stabilized fine-grained materials, however, the gap between the test error of a record-based split and the error obtained when an entire source is held out has not been quantified. The two kinds of split also differ in the source composition of the training set, the number of training records and the test records, and a direct comparison therefore confounds the effect of including records of the source being predicted (hereafter the target source) in training with these differences. To examine this effect, we designed a source-coverage comparison that fixes the number of training records and the test records and compares training sets with and without non-test records of the target source.

Besides strength itself, mixture design requires a model to predict how strength changes when the water content of the same material changes. A sensitivity analysis by Onyelowe et al. [9] showed that optimum water content had an effect on UCS close to zero. Reported water contents differ in meaning between studies, for example the optimum water content [5,7], the natural water content of the raw sludge [19] and the water content of the slurry before the stabilizer is added [2,20]. When water content in multi-source data varies mainly between studies, the water content–strength relationship learned by a model also reflects between-study differences in materials and specimen preparation. A matched water-content pair consists of two records of the same material that differ only in water content and gives the measured change in strength when water content changes. We used such pairs to test the direction and magnitude of the strength changes predicted by source models.

Once a few strength tests have been performed on a new material, the resulting records (hereafter local records) can be used to calibrate the predictions of a source model or to fit a local reference model directly. In material property prediction, Devi et al. [21] compared pretrained models fine-tuned on target data with models trained on the target data alone. The fine-tuned models had errors lower than or similar to those of the target-only models when 100–800 target records were

available, with 10 records as the exception. A few strength tests on a new material yield a number of records similar to this exception. We therefore used 10 local records to calibrate the source model and to fit local reference models, and compared the source model after calibration (hereafter the calibrated source model) with the local reference models on the same test records.

We trained source models on 399 UCS records of stabilized fine-grained materials from 10 sources and used four target materials absent from training to address three questions (Fig. 1). For new-material strength (Fig. 1a), we asked how much the error increased when a source model predicted a source absent from training and how target-source records in training affected the error when the number of training records and the test records were fixed. For water content (Fig. 1b), we asked how a water-content input affected the overall prediction error and whether models predicted the change in strength when the water content of the same material changed. For local records (Fig. 1c), we asked whether, with 10 local records, the calibrated source model had a lower error than the local reference models. The random forest was the main model, and five other models were used to check whether the results depended on model type. Public data on alkali-activated concrete (AAC) [22,23], which compile mixtures and strengths from multiple studies, were used to repeat the new-material strength and local-record comparisons on another class of material.

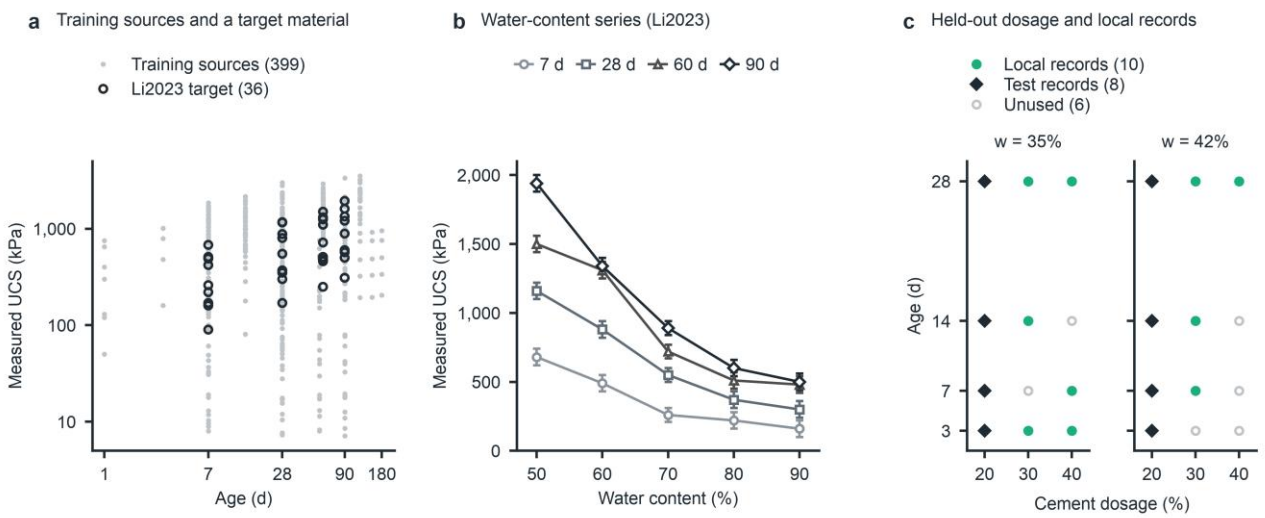


Fig. 1 Examples of the data used in the three comparisons. a, UCS versus curing age for the 399 records of the 10 training sources and the 36 records of the target material Li2023, on logarithmic axes. Li2023 was not used to train the source models. b, The 20 records of Li2023 at 20% cement dosage, covering 4 curing ages and 5 water contents. Lines connect records of the same age, and error bars show the reading tolerance (±50–60 kPa). The strength difference between two records of the same age is the ΔUCS of a matched water-content pair. c, One draw for Qiantang with the 20% dosage held out. The 8 test records are all records at 20% dosage, the 10 local records are drawn from the 30% and 40% dosages, and the remaining 6 records are unused. The calibrated source model and the local reference models use the same local records and predict the same test records.

2 Materials and methods

2.1 Datasets

Training data for the stabilized fine-grained materials were compiled from 33 publicly available documents, including research papers and technical reports. Of these, 22 contained extractable UCS records, 576 records in total. The materials were dredged sludge, engineering waste slurry and soft clays with similar properties, stabilized with cement, lime, fly ash or their combinations. After screening for material type, dosage basis, curing age and test conditions, 399 records were retained. The reasons for excluding the other 177 records are given in Supplementary Section S1 and Supplementary Table S1, and the use of each of the 33 documents is listed in Supplementary Table S2. Records from the same study and its associated documents were grouped into one source, as were different soil batches from the same study. The 399 records came from 10 sources (Table 1).

Table 1 Sources of the stabilized fine-grained materials and the target materials

Source	Ref.	Records	Water-content type	Water content (%)	Specimen preparation	Curing	Use
P01	[2]	40(0)	Initial slurry water content	100	Slurry-cast	Submerged	Training

Source	Ref.	Records	Water-content type	Water content (%)	Specimen preparation	Curing	Use
P02	[35]	3(0)	Prepared water content	70	Other or not reported	Sealed or wrapped	Training
P03	[36]	15(15)	Initial slurry water content	60	Other or not reported	Sealed or wrapped	Training
P04	[37]	8(8)	Initial slurry water content	109–138	Compacted	Sealed or wrapped	Training
P09	[7]	48(48)	Optimum water content	20.4–23.9	Compacted	Sealed or wrapped	Training
P15	[5]	120(0)	Optimum water content	18.5	Compacted	Sealed or wrapped	Training
P19	[38]	36(0)	Prepared water content	25	Compacted	Humidity-controlled	Training
P23	[39]	16(16)	Natural water content of raw sludge	100	Other or not reported	Humidity-controlled	Training
P25	[19]	25(0)	Natural water content of raw sludge	130	Slurry-cast	Sealed or wrapped	Training
P29 (Memphis, Mobile)	[20]	88(88)	Initial slurry water content	70–100	Slurry-cast	Humidity-controlled	Training
P37	[40]	72	Optimum water content	20.6–28.7	Compacted	Sealed or wrapped	Held-out dosages 3%, 5%, 7%, 10%
Li2023	[41]	36	Formulation water content	50–90	Slurry-cast	Humidity-controlled	Held-out dosages 10%, 15%, 20%, 25%, 30%, matched water-content pairs
Qiantang	[42]	24	Mixing water content	35–42	Slurry-cast	Sealed or wrapped	Held-out dosages 20%, 30%, 40%, matched water-content pairs
Zhaoqing	[43]	9	Mixing water content	60	Slurry-cast	Sealed or wrapped	Direct prediction

Table 1 note Numbers in parentheses after the record counts of the training sources are the numbers of records digitized from figures. P09, P15 and P37 did not state the mass basis of their optimum water contents. The water contents of Li2023, Qiantang and Zhaoqing are expressed relative to the mass of dry soil or dry sediment. Further details are given in Supplementary Sections S1 and S2. All four target materials were used for direct prediction, and the held-out dosages in the Use column were used in the comparisons of Section 2.5. Other inputs of the target materials are listed in Supplementary Table S4.

UCS, curing age and specimen dimensions were converted to kPa, d and mm, respectively, and stabilizer dosages were expressed as percentages of dry soil mass. Because water content was defined differently across studies, the reported values and their types were kept without conversion (Table 1). Clay content was taken as the reported fraction of particles finer than 2 μm. By this definition, clay content was available only for the 48 records from P09, all with a value of 14.5%. P25 reported a value of 79.8% without specifying the particle-size threshold, and this value was treated as missing in the main model. A sensitivity analysis that retains it is given in Supplementary

Section S4. For 175 of the 399 records (43.9%), strength was obtained by converting the coordinates of data points or bars in figures with the axis scales, with reading tolerances of 3–50 kPa. The extraction methods, the reading tolerance of each source and the basis for encoding the input variables are given in Supplementary Section S1 and Supplementary Table S3.

The four target materials were used to test source-model predictions for new materials, and none of their records came from the 33 documents above (Table 1, Supplementary Section S2). Each target material differed from the training materials, had model inputs traceable to the original publication and had extractable UCS records. In direct prediction, the source model trained on the 399 records predicted all records of each target material. P37 and Li2023 were used in preliminary analyses. The random forest configuration was fixed afterwards, and the Zhaoqing and Qiantang records were collected last. Fig. 2 compares the strength and water content of the training sources with those of the target materials.

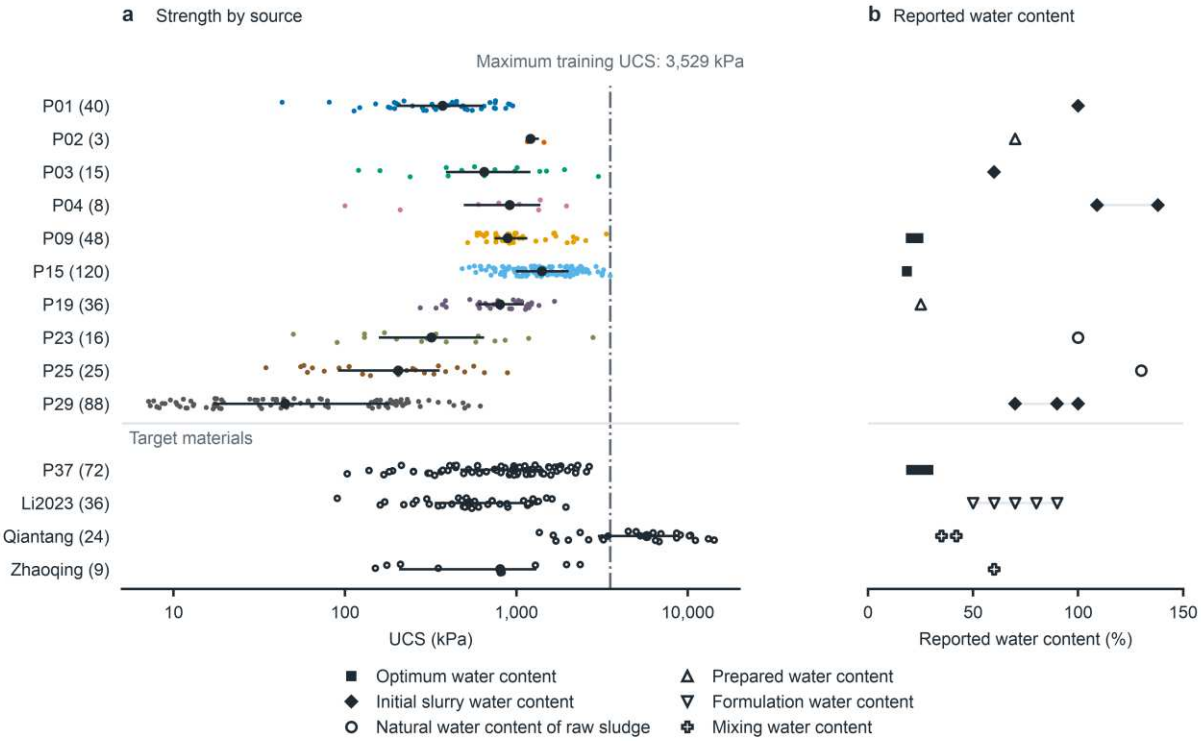


Fig. 2 Strength and water content of the training sources and target materials. a, The 399 records of the 10 training sources and the 141 records of the four target materials, on a logarithmic horizontal axis. Small points are individual records, colored by source for the training records and open for the target-material records. Large points mark the median of each source, and horizontal lines the interquartile range. Numbers in parentheses are record counts. The vertical dash–dot line marks the maximum UCS in the training data (3,529 kPa). b, Water content reported by each source. Horizontal lines connect the minimum and maximum values of each source, and symbols indicate the six water-content types in Table 1. The two panels are aligned by source.

The AAC data were taken from the public database of Moreno Torres et al. (Zenodo, version 2) [22,23]. Of its 1,630 records, the 700 with a positive 28-day cube compressive strength were selected. The database assigns a reference number (Ref) to each original publication, and the 700 records belonged to 52 Refs, each treated as one source. Within a Ref, records with identical mixture proportions that differed only in curing conditions or cube size formed a mixture group. Data were split by mixture group, and all records of a group were held out together. The nine Refs with at least 20 records and at least 10 mixture groups served as target sources, giving 349 records for the source-coverage comparison and the calibration comparison (Supplementary Table S5). Because Ref51 and Ref76, and Ref77 and Ref89, reported results for the same mixtures [24–27], the other Ref of a pair was excluded from the training set whenever one of them was predicted. The input values of these four Refs were checked against the original publications, and 235 input values

for mixture proportions and curing conditions were corrected, whereas the strength values were left unchanged (Supplementary Section S3).

2.2 Input schemes and source models

Three input schemes (Base, Water and Full) were used for the stabilized fine-grained materials, with variable definitions given in Supplementary Table S6. The Base scheme comprised the cement, lime and fly ash dosages and the curing age, entered as ln(1 + age in days). The Water scheme added water content to the Base scheme, and the Full scheme added plasticity index, clay content, organic matter content, specimen aspect ratio, preparation category and curing category to the Water scheme. We added each of the seven additional Full-scheme variables to the Base scheme in turn and compared the random forest's source-averaged mean absolute error (MAE) under LOSO validation (Section 2.6). Water content gave the largest reduction and was therefore selected for the Water scheme (Section 3.2.1). Source and record identifiers were not used as inputs.

In the Full scheme, clay content, organic matter content and specimen aspect ratio had missing values (Supplementary Table S6). Missing values were imputed with the training-set median, and each variable with missing values was paired with an indicator marking whether its value was missing. Numerical variables were standardized and categorical variables one-hot encoded, with all imputation, standardization and encoding parameters estimated from the training records.

Clay content was available only for records from P09 (Section 2.1). Whenever P09 was in the training set, median imputation made clay content constant, and its missing-value indicator, which coincided with whether a record came from P09, acted as a source identifier for P09. To examine the effect of this identifier, a sensitivity analysis removed clay content and its missing-value indicator, kept all other settings unchanged and repeated the main comparisons with the same data splits, local-record draws and source resampling (Supplementary Section S4, Supplementary Table S8).

For the AAC, the inputs were 14 mixture variables, four curing variables, cube side length and final curing state. Blank database entries were treated as missing values and preprocessed as for the stabilized fine-grained materials (Supplementary Section S3).

The source models were a random forest [28], the main model, and five comparison models: ridge regression (linear model), extra trees and XGBoost [29] (tree ensembles), Gaussian process regression [30] (kernel method) and a multilayer perceptron (neural network). All six models were trained on $\ln(1 + \mathrm{UCS/kPa})$, and a prediction z was back-transformed to strength as $\exp(z) - 1$, with negative values set to zero. Except in Gaussian process regression, each training record was weighted by the reciprocal of the number of training records from its source, giving every source the same total weight. Because the Gaussian process regression implementation did not support sample weights, all records had the same weight in that model (Supplementary Section S4).

The random forest had 500 trees, considered a random 80% of the input variables at each split and required at least two records per leaf. This configuration was fixed after a preliminary comparison of ridge regression, random forest, extra trees and gradient-boosted trees and was kept unchanged in all later analyses (Supplementary Table S7). The random forest for the AAC used the same settings, was trained on strengths converted from MPa to kPa and gave each Ref the same total weight.

The comparison models used the same Full scheme as the random forest and were evaluated with the record random split and LOSO validation (Section 2.3) to test whether the conclusions depended on model type. The multilayer perceptron had two hidden layers and was trained with five random seeds, and the log-strength predictions of the five networks were averaged. For XGBoost, the sample weights were normalized to a mean of 1 and the initial prediction was set to the median log strength of the training set. These settings were adopted after the initial configuration produced extreme predictions. Model parameters are given in Supplementary Section S4.

2.3 Validation splits and source-coverage comparison

The data for the stabilized fine-grained materials had two levels of grouping. Records from the same source shared raw materials and specimen preparation methods, and within a source, the

strengths of one mixture measured at different curing ages formed a mixture group. The 399 records formed 105 mixture groups in 10 sources (grouping criteria in Supplementary Section S5). The three validation splits differed in the unit held out for testing (Supplementary Fig. S1). The record random split held out a random 20% of individual records, regardless of their mixture group or source. The mixture-group split held out a random 20% of the mixture groups, keeping all records of a group in either the training set or the test set. LOSO validation held out all records of one source at a time and trained the models on the remaining sources. The record random split and the mixture-group split were each repeated 20 times, and LOSO validation comprised 10 folds, with the two material batches of P29 held out together as one source. The random forest was evaluated under all three splits and the five comparison models under the record random split and LOSO validation. Because each P19 mixture was molded at four dry densities, P19 was also grouped by dry density, giving 114 mixture groups, and the mixture-group split was repeated with this grouping (Supplementary Section S5, Supplementary Table S9).

Relative to the record random split, LOSO validation removed test-source records from the training set and also changed the number of training records and the test records (Section 1). To compare errors with and without test-source records in training, we designed a source-coverage comparison (Supplementary Fig. S1) in which two training sets of equal size predicted the same test records. The target-covered group replaced m records from other sources in the target-uncovered group with m non-test records from the target source, and at least one record from each other source remained after the replacement.

In each comparison, a random 20% of the target-source records (rounded up) were taken as test records, and the remaining m records were used for training. The training set of the target-uncovered group consisted of 279 records from the other sources, with at least one record per source. The number 279 was the total number of records outside the target source when P15 was the target, the smallest such total among the 10 sources. The same model was trained on each training set, both fitted models predicted the same test records, and the MAEs of the two groups were

compared. Each of the 10 training sources served in turn as the target source, with 50 random draws per target. The AAC comparison followed the same procedure with 490 training records per set, m = 10 and 50 repeats per target Ref, and its test records and 10 target-Ref records were those of the calibration comparison (Section 2.5). All six source models used the same draws, and sampling details are given in Supplementary Section S5.

Two supplementary checks examined the effects of training weights and of the size of the largest sources. In the main analysis, each training record was weighted by the reciprocal of the number of training records from its source, giving all sources the same total weight (Section 2.2). The weighting check repeated the source-coverage comparison with every record given the same weight in the random forest, under which sources with more records carried more weight. The record-count check drew 40 records from each of the two largest sources, P15 (120 records) and P29 (88 records), with the P15 records drawn at random and the P29 records stratified by material batch (22 from Memphis and 18 from Mobile). With the other sources unchanged, this gave 20 subsets of 271 records each, and the split error ratio and the source-coverage error ratio of the random forest (Section 2.6) were recalculated for each subset (Supplementary Section S5).

2.4 Water-content input and matched pairs

The random forest was used to compare the overall prediction errors of the Water and Full schemes. Under LOSO validation, the MAEs of the two schemes were compared source by source for the 10 training sources, and both schemes were also used to predict Zhaoqing and Qiantang directly (Section 2.1). The comparison in which each variable was added to the Base scheme is described in Section 2.2. To quantify how the reported water contents varied between and within sources, the total sum of squares of water content over the 399 records was decomposed into between-source and within-source components (Supplementary Section S6).

Strength changes following a change in water content were evaluated with matched water-content pairs. A pair consisted of two records from the same batch of material that had the same soil, cement type, cement dosage and curing age and differed only in water content (Fig. 1b). The 68

pairs came from four batches of material. Memphis and Mobile, two batches of P29, provided 24 pairs (water contents of 90% and 100%) and 16 pairs (70% and 100%), respectively, and Qiantang provided 12 pairs (35% and 42%). Li2023 had five records at 20% cement dosage at each of four curing ages (7, 28, 60 and 90 d), with water contents of 50%, 60%, 70%, 80% and 90%. At each age, records with adjacent water contents formed pairs, giving four pairs per age and 16 in total. Adjacent pairs shared a record, so these 16 pairs were not independent. Pairing all five records at the same age gave 10 pairs per age and 40 in total, including the 16 adjacent pairs.

For each pair, the measured ΔUCS was the UCS at the higher water content minus that at the lower water content, and the predicted ΔUCS was calculated in the same way from the predictions of the same model for the two records. Predicted and measured changes were compared as described in Section 2.6. The reading tolerance of a pair was the sum of the tolerances of its two records, which gave ±6, ±10, ±100 and ±110–120 kPa for Memphis, Mobile, Qiantang and Li2023, respectively.

Because Memphis and Mobile belong to P29, they were predicted by models trained without P29, whereas Qiantang and Li2023, both target materials, were predicted by models trained on all 399 records. The Memphis, Mobile and Qiantang pairs were predicted by the six source models under the Water and Full schemes (Supplementary Table S13). The 16 adjacent Li2023 pairs were predicted by the random forest, ridge regression, Gaussian process regression and multilayer perceptron under the Full scheme, and the 40 all-combination pairs by the random forest under the Water and Full schemes (Supplementary Section S6).

Each tree in a random forest gives the same prediction for all records in a leaf node, so a tree predicts zero change for a pair whose two records fall into the same leaf. To interpret the random forest results for the pairs, we calculated the proportion of trees in which the two records of a pair fell into the same leaf node and recorded the water-content split points of the trees. This analysis was added after the pair results had been examined (Supplementary Section S6).

2.5 Calibration of source models and comparison with local reference models

We assumed that a new material had been tested at only a few mixtures. Its local records were used to predict the strength of its untested mixtures, either by calibrating the source-model predictions with the local records or by fitting a local reference model to the local records alone. The calibrated source models and the local reference models used the same local records and predicted the same test records, and their prediction errors were compared (hereafter the calibration comparison, Table 2).

Each target material had been tested at several cement dosages. In each test task, one dosage was treated as untested, all records at that dosage formed the test set, and 10 local records were drawn at random, without replacement, from the records at the other dosages. The draw was repeated 200 times per task, giving a set of 10 local records each time (Fig. 1c). P37, Li2023 and Qiantang had 4, 5 and 3 test tasks, respectively, and the cement dosage of each task is listed in Table 1. Analyses with five local records used the first five records of the corresponding 10-record draw (Supplementary Section S7).

The source model was a Full-scheme random forest trained on all 399 training records. Three calibration methods (mean shift, median shift and log-linear calibration) adjusted the source-model predictions with the local records, and the uncalibrated source model served as the baseline (Table 2). The local reference models were fitted to the local records alone and comprised the local median, a log-scale local ridge regression and a raw-scale local ridge regression (Table 2). The raw-scale local ridge regression was added as a post hoc sensitivity analysis after the log-scale fit produced extreme predictions. For log-linear calibration and both local ridge regressions, the standardization parameters were calculated from the local records, and all records had the same weight in fitting. Negative predictions were set to zero for all methods, and predictions had no upper bound (Supplementary Section S7).

For Qiantang, calibrated random forests with the Base and Water schemes were compared alongside the Full scheme, and the evaluation also covered the ΔUCS prediction error for the four water-

content pairs within each held-out dosage (Section 2.4). To test whether the results depended on the type of source model, the calibration comparison for P37, Li2023 and Qiantang was repeated with Gaussian process regression as the source model, using the same local records and test records as for the random forest (Supplementary Section S7).

For the AAC, the calibration comparison was repeated on the nine target Refs. In each draw, 20% of the mixture groups of the target Ref (rounded up) were held out as test records, and 10 local records were drawn at random, without replacement, from the remaining mixture groups of the same Ref. For each target Ref, 200 splits were used, each with at least five test records and at least 10 remaining records. The training set of the source model excluded the target Ref and its associated Ref (Section 2.1).

Table 2 Source-model calibration methods and local reference models

Category	Method	Input	Fitting target	Prediction
Baseline	Uncalibrated source model	Full scheme	—	Direct prediction by the source model
Source-model calibration	Mean shift	Measured UCS minus source-model predictions for the local records	—	Source-model prediction plus the mean of the differences
	Median shift	Same as above	—	Source-model prediction plus the median of the differences
	Log-linear calibration	ln(1 + source-model prediction/kPa)	ln(1 + UCS/kPa)	Ridge regression with intercept, α = 1
Local reference	Local median	—	—	Median UCS of the local records
	Log-scale local ridge regression	Cement dosage, ln(1 + age/d) and water content, plus fly ash dosage for P37	ln(1 + UCS/kPa)	Ridge regression, α = 1
	Raw-scale local ridge regression	Same as above	UCS	Ridge regression, α = 1, post hoc sensitivity analysis

Table 2 note The local ridge regression for the AAC used the same inputs as the source model.

2.6 Error summary and uncertainty

Prediction error was measured by the MAE (Eq. 1).

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \qquad (1)$$

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \qquad (2)$$

Here, n is the number of test records, and y_i and $\hat{y}_i$ are the measured and predicted UCS of the ith record, in kPa for the stabilized fine-grained materials and in MPa for the AAC. For the record random split and LOSO validation, the coefficient of determination R^2 was calculated on the original strength scale by Eq. (2), where $\bar{y}$ is the mean measured UCS of the test records. The R^2 of the record random split was the median over the 20 splits.

Because the training sources contained 3–120 records each (Table 1), prediction error was summarized in two ways. The overall MAE pooled all test records in Eq. (1) and thus weighted sources by their numbers of records. The source-averaged MAE was the mean of the per-source MAEs from Eq. (1) over the S sources (Eq. 3), giving every source the same weight. For the AAC, the sources were the Refs.

$$\mathrm{MAE}_{SA} = \frac{1}{S}\sum_{s=1}^{S} \mathrm{MAE}_s \qquad (3)$$

In Eq. (3), MAE_s is the MAE of the sth source.

The split comparison and the source-coverage comparison were each summarized by an error ratio. The record random split and the mixture-group split were each repeated 20 times, and the median overall MAE of the 20 repeats was used for each. For LOSO validation, the overall MAE was calculated from the 399 pooled predictions. The split error ratio was the overall MAE of LOSO validation divided by the median overall MAE of the record random split (Eq. 4), whereas the error of the mixture-group split was reported separately and not included in this ratio. In the source-coverage comparison, the MAE of each target source was first averaged over its 50 repeats, and the source-averaged MAE of each group was then calculated by Eq. (3). The source-coverage error ratio was calculated by Eq. (5), and a value above 1 indicates a lower error when the training set contains records of the target source.

$$R_{\mathrm{split}} = \frac{\mathrm{MAE}_{\mathrm{LOSO}}}{\mathrm{median}\left(\mathrm{MAE}_{\mathrm{RR},1}, \ldots, \mathrm{MAE}_{\mathrm{RR},20}\right)} \qquad (4)$$

$$R_{\mathrm{cov}} = \frac{\mathrm{MAE}_{SA,\mathrm{uncovered}}}{\mathrm{MAE}_{SA,\mathrm{covered}}} \qquad (5)$$

For the water-content pairs, the predicted ΔUCS was compared with the measured ΔUCS. A predicted direction was counted as consistent when the predicted and measured ΔUCS had the same sign. Only pairs whose absolute measured ΔUCS exceeded the pair reading tolerance (Section 2.4) entered the direction count, and a predicted ΔUCS of zero was counted as inconsistent. The magnitude of the predicted change was assessed by comparing the medians of the predicted and measured ΔUCS. The absolute difference between the predicted and measured ΔUCS of a pair was taken as the absolute ΔUCS error, and its median over the pairs was reported (Supplementary Section S6).

For each draw, we calculated the MAE difference and ratio between the calibrated source model and the local reference model on the common test records (Eq. 6), and reported the median and interquartile range over the 200 draws. For P37 and Li2023, which had several test tasks, the task MAEs were first averaged with equal weights within each draw. Qiantang was reported separately for each dosage. Each held-out dosage of Qiantang contained four water-content pairs, and in each draw the mean absolute error of the predicted ΔUCS of these pairs was calculated, with the median over the 200 draws reported. For the AAC, the 200 MAEs of each Ref were averaged first, the nine Refs were then averaged with equal weights by Eq. (3), and the error ratio was the ratio of the source-averaged MAEs of the two methods. All extreme predictions were retained in the summaries.

$$\Delta\mathrm{MAE} = \mathrm{MAE}_{\mathrm{cal}} - \mathrm{MAE}_{\mathrm{loc}}, \qquad r = \frac{\mathrm{MAE}_{\mathrm{cal}}}{\mathrm{MAE}_{\mathrm{loc}}} \qquad (6)$$

In Eq. (6), $\mathrm{MAE}_{\mathrm{cal}}$ and $\mathrm{MAE}_{\mathrm{loc}}$ are the MAEs of the calibrated source model and the local reference model, respectively.

The 95% intervals describe how the results varied with the composition of the sources. Because records from the same source shared raw materials and specimen preparation methods, the intervals were estimated by resampling sources rather than records. In each resample, S sources were drawn

with replacement from the S sources, a source drawn more than once was counted each time, and the summary error, difference or ratio was recalculated. This was repeated 5,000 times, and the 2.5th and 97.5th percentiles formed the 95% interval. For the difference between the calibrated source model and the raw-scale local ridge regression on the AAC, the lower bound varied with the random seed, so the interval was instead obtained by enumerating all 24,310 resampling combinations of the nine Refs. Each combination was weighted by its multinomial probability, which is equivalent to the 9^9 equally likely ordered resamples, and the 2.5th and 97.5th percentiles of this distribution were taken (Supplementary Section S7). The two groups or two methods in a comparison used the same resamples, and the numerator and denominator of the split error ratio were recalculated in each resample (Supplementary Section S5). The models were not retrained during resampling, and their predictions were held fixed.

To test the influence of individual sources, the error ratios were recalculated with each source removed in turn. For the split comparison and the source-coverage comparison, a sign test counted the signs of the source-level differences in error between the two conditions, and a one-sided exact test assessed whether the number of sources with a lower error under one condition exceeded the number expected if the two conditions did not differ (Supplementary Section S5).

Analyses used Python 3.13.13, scikit-learn 1.9.1 [31] and XGBoost 3.4.1, and the full software environment is documented with the code. The scripts for data extraction, model analysis and plotting were written with the assistance of OpenAI Codex, and the authors take responsibility for the submitted code, results and text.

3 Results

3.1 Effects of the validation split and training-set source composition on error

3.1.1 Errors under the three validation splits

The median R^2 over the 20 record random splits was 0.91 (0.92 for the split shown in Fig. 3a), whereas the 399 LOSO predictions gave an R^2 of 0.12 (Fig. 3b). The overall MAE of the random forest was 133.9 kPa under the record random split and 200.4 kPa under the mixture-group split

406 (medians of 20 splits each), rising to 500.6 kPa under LOSO validation (Fig. 3c, Supplementary
407 Table S9). The split error ratio was 3.74 (95% interval, 2.26–4.59), and the LOSO MAE exceeded
408 the record-random-split MAE in all 10 sources (Supplementary Fig. S2, Supplementary Section
409 S5).

410 Across the six source models, the split error ratio ranged from 2.46 to 6.52, and Gaussian process
411 regression had the lowest overall LOSO MAE (Fig. 3d, Supplementary Table S10). The median
412 record-random-split MAE of the multilayer perceptron was 44.2 kPa lower than that of the random
413 forest (95% interval, −81.5 to −23.2 kPa). Under LOSO validation, the MAE of the multilayer
414 perceptron exceeded that of the random forest by 83.8 kPa, with a 95% interval of −108.5 to 241.6
415 kPa that included zero (Supplementary Section S5).

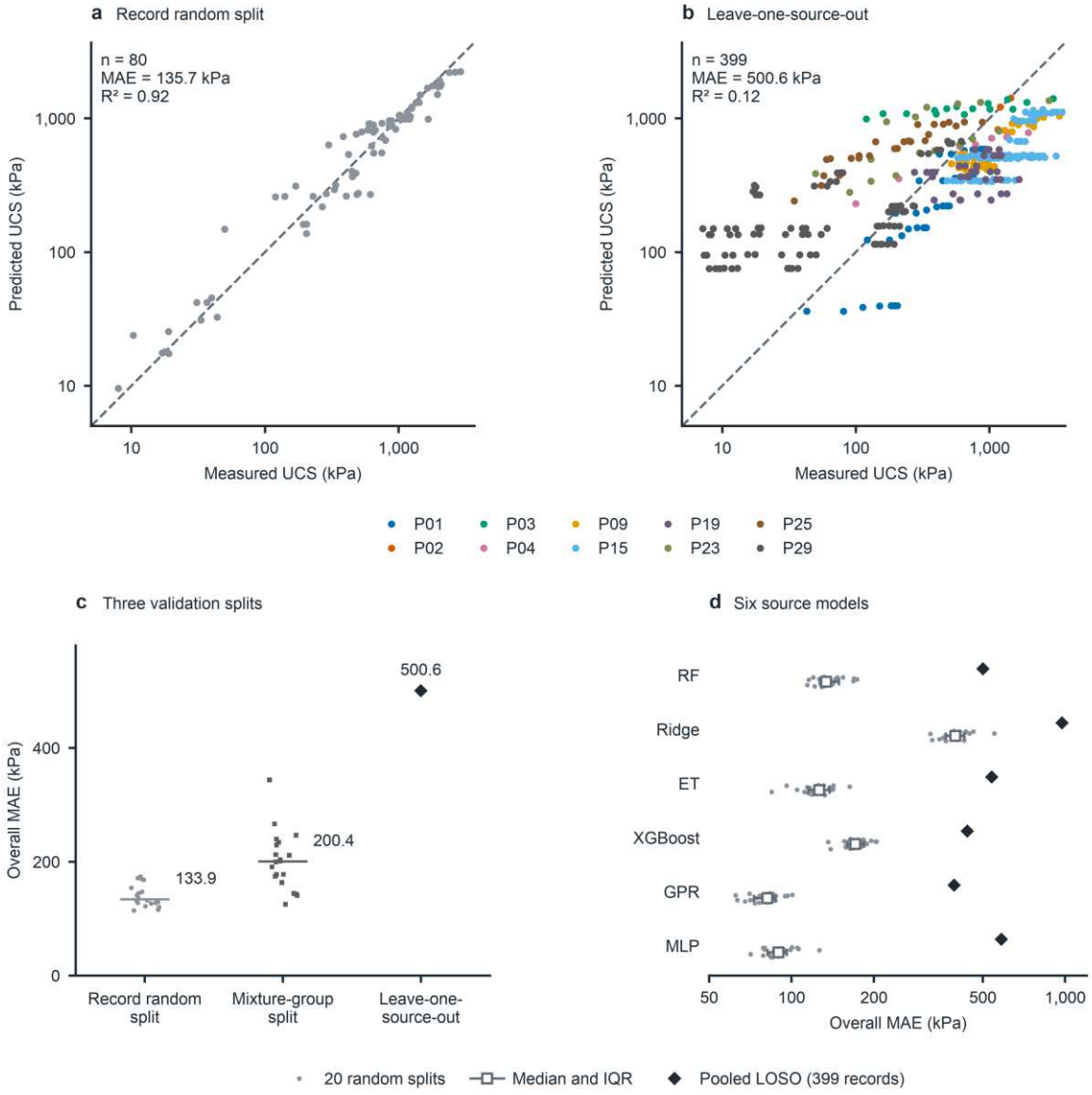


Fig. 3 Prediction errors under the three splits. a, Test predictions (n = 80) of the Full-scheme random forest from the record random split whose MAE was closest to the median of the 20 splits. b, The 399 LOSO predictions of the same model, colored by source. Panels a and b share logarithmic axes, the dashed line is the 1:1 line, and MAE and R^2 were calculated on the original strength scale. c, Overall MAE of 20 record random splits and 20 mixture-group splits. Small points are individual splits, horizontal lines and numbers are medians, and the diamond is the overall MAE of the 399 pooled LOSO predictions. d, Overall MAE of the six source models under the record random split and LOSO validation. Light gray points are the 20 random splits, open squares with horizontal lines are medians and interquartile ranges, and dark diamonds are the overall MAE of the 399 pooled LOSO predictions. The horizontal axis is logarithmic. RF, Ridge, ET, XGBoost, GPR and MLP denote random forest, ridge regression, extra trees, XGBoost, Gaussian process regression and multilayer perceptron, respectively, and the model configurations are described in Section 2.2.

3.1.2 Effect of target-source records on error

In the source-coverage comparison (Section 2.3), errors were lower when the training set contained records of the target source (Fig. 4). For the random forest, the source-coverage error ratio was 2.57 (95% interval, 1.89–3.69), and the target-covered group had the lower error for all 10 target sources.

The corresponding ratio for AAC was 1.77 (95% interval, 1.46–2.06). Across the six source models on the stabilized fine-grained materials, the ratio ranged from 2.09 to 5.79, and the lower limits of all 95% intervals exceeded 1 (Fig. 4c, Supplementary Fig. S3). In the four checks, the random forest's source-coverage error ratio was 2.55 with equal record weights, 2.26–2.76 when one source was removed at a time, 2.34–2.54 across the 20 subsets formed by downsampling P15 and P29, and 2.60 without clay content. In the three checks other than equal-record-weight training, the split error ratio ranged from 2.61 to 4.03 (Supplementary Section S5, Supplementary Table S8, Supplementary Fig. S3).

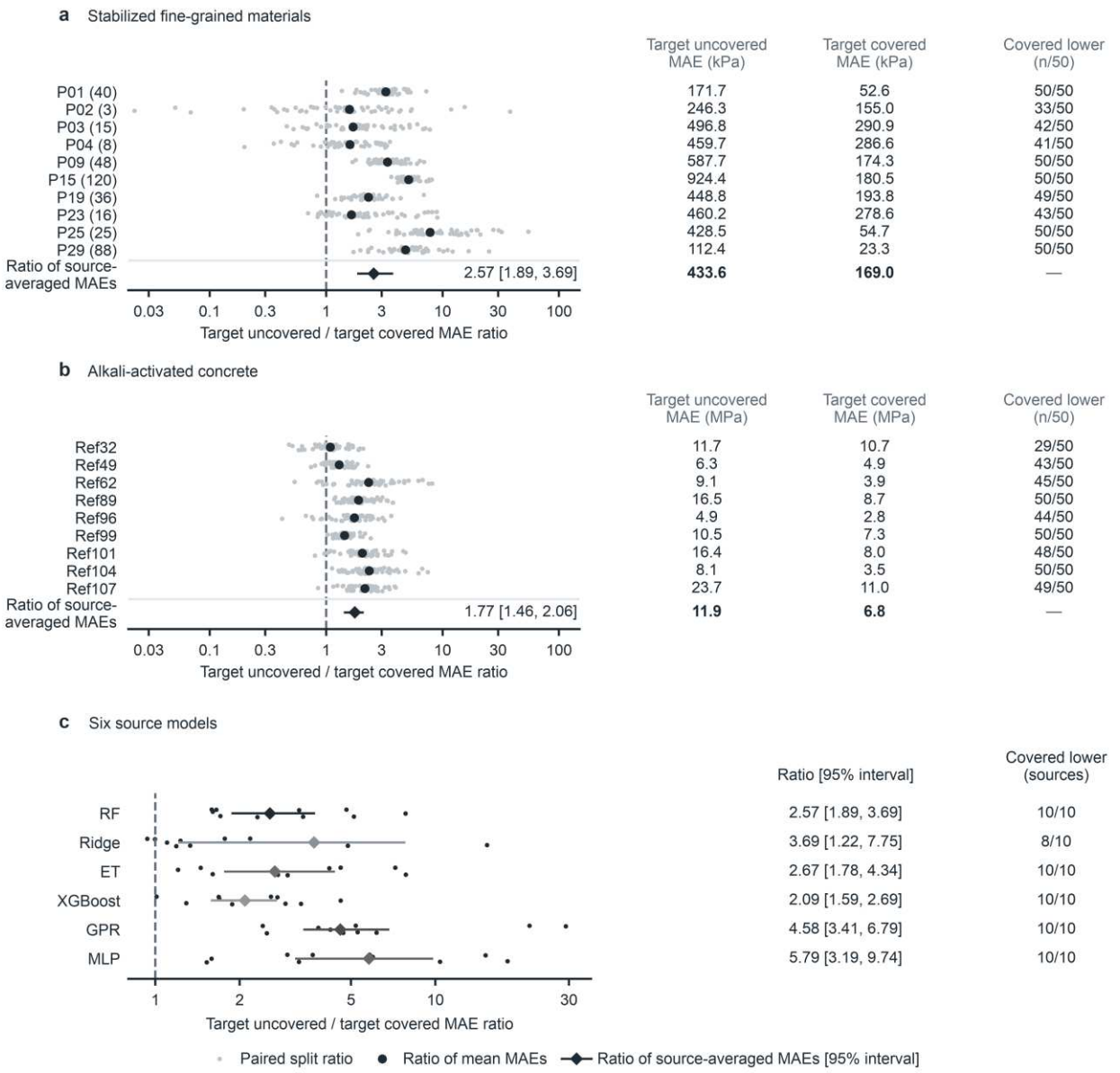


Fig. 4 Source-coverage comparison for the two material types. a, The 10 sources of the stabilized fine-grained materials, with record counts in parentheses. b, The 9 Refs of the alkali-activated concrete (AAC). For each target source, 50 draws were made. Within each draw, the two groups used the same test records and the same number of training records. Small points are the single-repeat ratios of the target-uncovered MAE to the target-covered MAE, and large dots are the ratios of the 50-repeat mean MAEs of the two groups. The columns on the right give the mean MAE of each group and the number of repeats in which the target-covered group had the lower error, in kPa for a and in MPa for b. The diamond at the bottom is the ratio of the source-averaged MAEs of the two groups. c, Comparison of the six source models. In each row, small points are the ratios of the 50-repeat mean MAEs for the 10 individual sources, and the diamond is the ratio of the source-averaged MAEs of the two groups. The columns on the right give the ratio with its 95% interval and the number of sources in which the target-covered group had the lower error. The summary segments in a–c are 95% intervals from 5,000 paired source resamples with the predictions fixed. The vertical dashed line marks a ratio of 1.

3.2 Effects of water-content input on predicted strength and strength change

3.2.1 Strength prediction error

When each of the seven variables was added to the Base scheme in turn, water content gave the lowest source-averaged MAE, reducing it from 679.2 to 362.1 kPa (Fig. 5a). The source-averaged MAE of the Water scheme was 13.1% lower than that of the Full scheme, and the Water scheme had the lower MAE in 8 of the 10 sources (Fig. 5b). In direct prediction, the Water scheme had a lower MAE than the Full scheme for Qiantang and a higher MAE for Zhaoqing (Supplementary Fig. S4 and Supplementary Table S12).

In the training data, 7 of the 10 sources had a single water-content value, and between-source differences accounted for 97.75% of the total sum of squares of water content (Fig. 2b, Supplementary Section S6). Of the 24 Qiantang records, 15 had a UCS above the training maximum of 3,529 kPa. For predictions bounded above by 3,529 kPa, the lower bound of the MAE for Qiantang was 3,064.4 kPa (Fig. 2a, Supplementary Section S6).

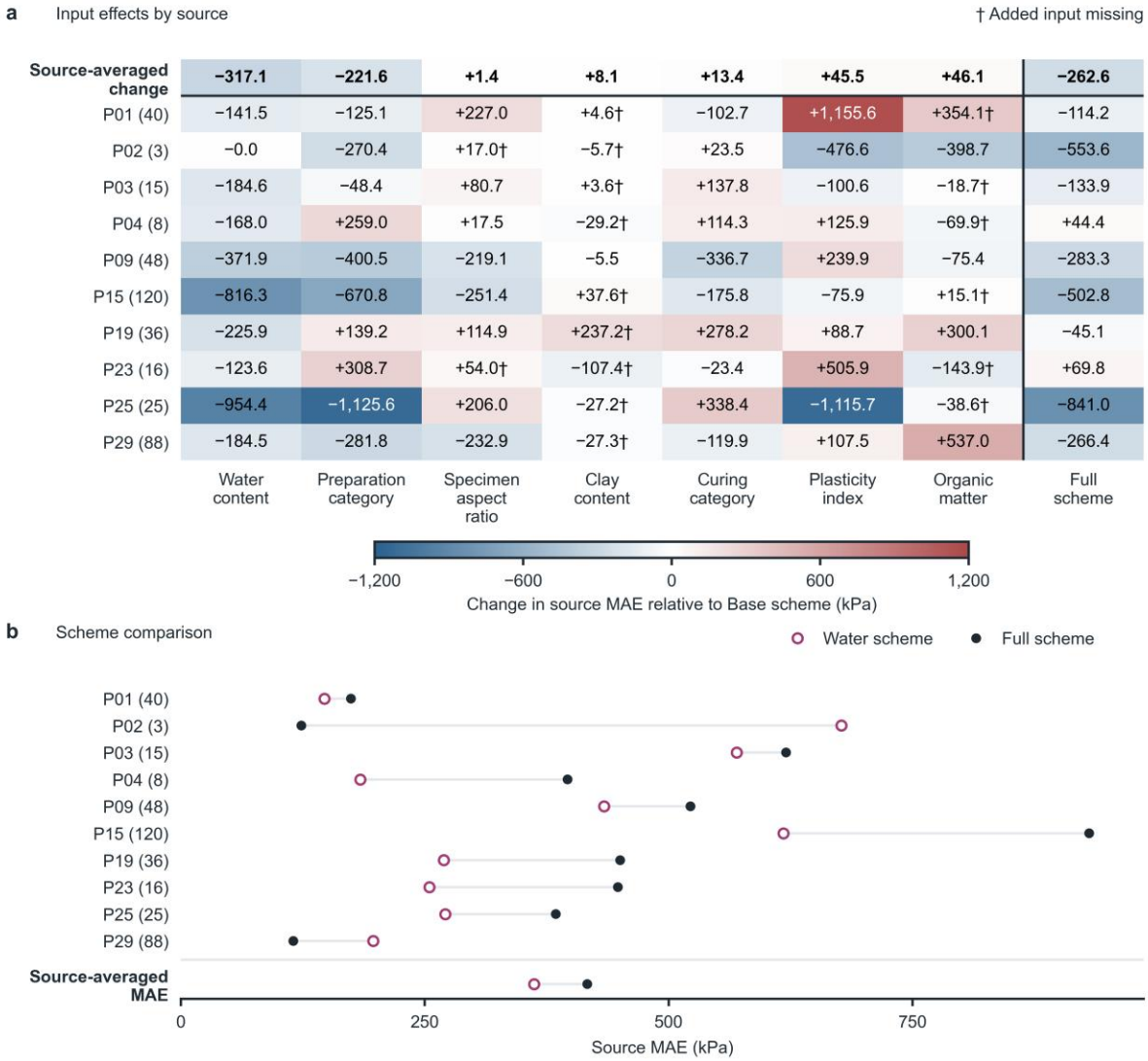


Fig. 5 Input schemes and LOSO prediction errors. a, Change in the MAE of each source (scheme minus Base scheme, in kPa) when each of the seven variables was added to the Base scheme, and for the Full scheme. Negative values indicate a lower error. The first row is the arithmetic mean over the 10 sources, the first seven columns are ordered by this mean from lowest to highest, and the last column is the Full scheme. Numbers in parentheses are record counts. A dagger (†) marks a variable that was missing from the model input for all records of that source before preprocessing. b, MAE of each source and source-averaged MAE for the Water and Full schemes. Both panels are based on the random forest and the 399 LOSO predictions, and the values are listed in Supplementary Table S11.

3.2.2 Prediction of strength change within the same material

The median measured ΔUCS was −8.7, −106.5, −1,770.0 and −200.0 kPa for Memphis, Mobile, Qiantang and Li2023, respectively (Fig. 6, Supplementary Table S13, Supplementary Section S6). Under the Full scheme, the random forest predicted zero change for Memphis and a median change of −0.5 kPa for Qiantang, and XGBoost predicted zero change for both batches. Among pairs whose measured |ΔUCS| exceeded the pair reading tolerance, ridge regression and the multilayer

perceptron predicted the measured direction in all four batches, whereas Gaussian process regression did so for 7 of the 12 Qiantang pairs. For Qiantang, the predicted decreases were smaller than the measured decrease. The predicted decrease exceeded the measured decrease for ridge regression, Gaussian process regression and the multilayer perceptron on Memphis, for the random forest and Gaussian process regression on Mobile, and for Gaussian process regression and the multilayer perceptron on Li2023.

For every Memphis pair, the two records fell into the same leaf node in all 500 trees of the random forest, so the predicted change was zero. Both Memphis water contents, 90% and 100%, lay on the same side of the main split point at 85%, whereas the two Mobile water contents, 70% and 100%, lay on opposite sides (Supplementary Section S6).

For Memphis, the random forest and XGBoost gave a median predicted ΔUCS of zero under both schemes. For the other four models on Memphis and all six models on Mobile, the Water scheme predicted a larger decrease than the Full scheme. For 5 of the 6 models on Mobile, the median absolute ΔUCS error was higher under the Water scheme than under the Full scheme (Supplementary Table S13). This included the random forest, whose overall LOSO MAE was lower with the Water scheme than with the Full scheme (Section 3.2.1).

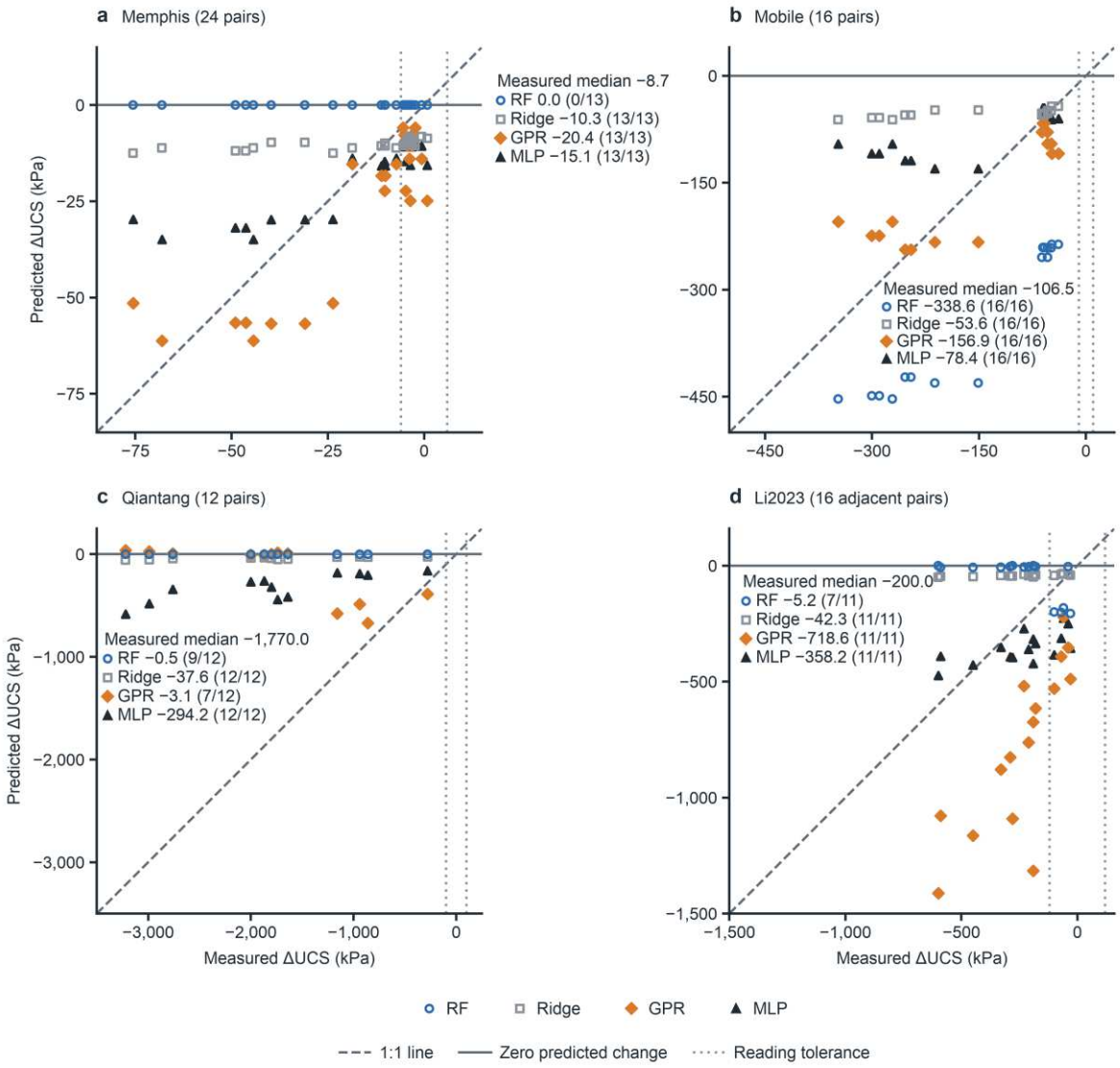


Fig. 6 Measured and predicted ΔUCS of matched water-content pairs for four batches of material. a, Memphis, 24 pairs, water contents of 90% and 100%. b, Mobile, 16 pairs, 70% and 100%. c, Qiantang, 12 pairs, 35% and 42%. d, Li2023, 16 pairs of adjacent water contents between 50% and 90% at the same curing age. The horizontal axis is the measured ΔUCS and the vertical axis the ΔUCS predicted by four source models under the Full scheme, both calculated as the UCS at the higher water content minus that at the lower water content (kPa). The two axes of each panel share the same scale, which differs between panels. The models for Memphis and Mobile were trained with P29 held out, and those for Qiantang and Li2023 on all 399 records. The dashed diagonal is the 1:1 line, and the horizontal solid line marks zero predicted change. Vertical dotted lines mark the pair reading tolerances of ±6, ±10 and ±100 kPa in a–c. In d, the reading tolerance of each pair is ±110 or ±120 kPa, and the lines are drawn at ±120 kPa. Each panel gives the median predicted ΔUCS of the four models, and the numbers in parentheses are the number of pairs whose predicted direction matched the measured one and the number of pairs whose measured |ΔUCS| exceeded the pair reading tolerance. Results of all six models under both input schemes for Memphis, Mobile and Qiantang are given in Supplementary Table S13, and the pair statistics for Li2023 in Supplementary Section S6.

3.3 Prediction error after local records become available

3.3.1 Errors across target materials

With 10 local records and log-linear calibration, the random forest had a lower median MAE than all three local reference models for P37 and Li2023, but a higher median MAE than both local ridge regressions in all three Qiantang dosage tasks (Fig. 7a, Supplementary Tables S14 and S15). Relative to the raw-scale local ridge regression, its median paired MAE ratios were 0.93 for P37 and 0.80 for Li2023, and the interquartile range for P37 included 1 (Supplementary Fig. S5). The calibrated Gaussian process regression had a higher error than the raw-scale local ridge regression for all three target materials, with median paired ratios of 1.07 for P37 and 2.27 for Li2023, and the interquartile range for P37 included 1 (Supplementary Table S19).

Compared with the uncalibrated random forest, log-linear calibration lowered the median MAE for P37, Li2023 and the 30% and 40% Qiantang dosage tasks and raised it for the 20% Qiantang task (Fig. 7a, Supplementary Table S14). In the 20% Qiantang task, the calibrated Base-scheme random forest had a median MAE of 789.8 kPa. Its median paired MAE difference from the raw-scale local ridge regression was −64.3 kPa, with an interquartile range of −258.3 to 419.3 kPa that included zero. The Base scheme predicted zero change for the four matched water-content pairs, and the mean absolute error of its predicted ΔUCS was 810.0 kPa (Fig. 7c, Supplementary Section S7).

For AAC with 10 local records, the calibrated random forest had a lower MAE than the local median, with a difference (calibrated minus local) of −0.46 MPa (95% interval, −0.90 to −0.07), and a higher MAE than the raw-scale local ridge regression, with a difference of +3.53 MPa (95% interval, 0.044 to 7.60). With one Ref removed at a time, the mean MAE difference from the raw-scale local ridge regression ranged from 2.14 to 4.32 MPa, and the interval included zero in 5 of the 9 analyses (Fig. 7b, Supplementary Tables S16 and S17, Supplementary Section S7).

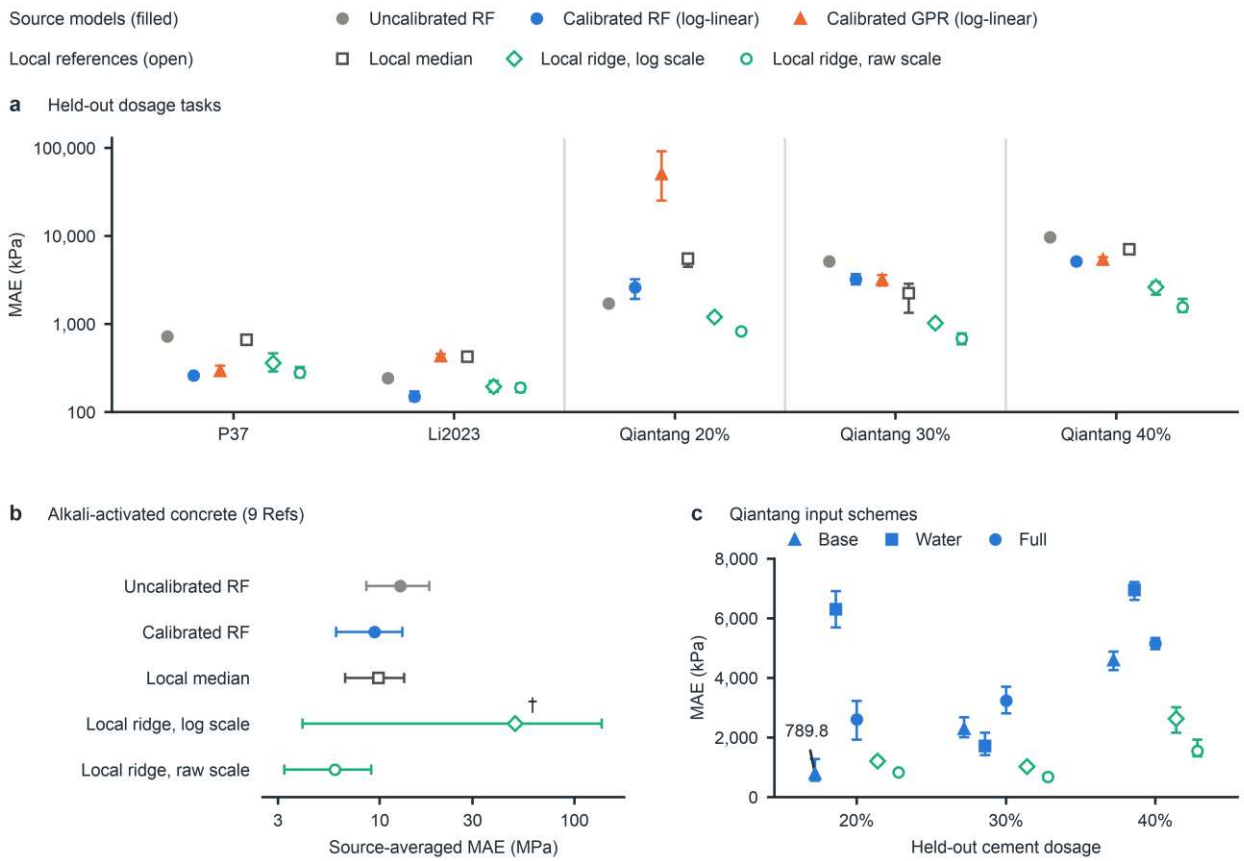


Fig. 7 Errors of the calibrated source models and local reference models with 10 local records. All methods used the same local records and test records, and calibration refers to log-linear calibration. a, MAE of the held-out dosage tasks for P37, Li2023 and Qiantang, on a logarithmic vertical axis. Filled symbols denote the uncalibrated random forest, the calibrated random forest and the calibrated Gaussian process regression, and open symbols the local median, the log-scale local ridge regression and the raw-scale local ridge regression. The source models used the Full scheme. For P37 and Li2023, the MAEs of the 4 and 5 dosage tasks, respectively, were averaged with equal weights within each draw. Points and error bars are medians and interquartile ranges over the 200 draws, and the predictions of the uncalibrated random forest do not depend on the local draw. b, Source-averaged MAE of the 9 AAC Refs, on a logarithmic horizontal axis. Error bars are 95% intervals from 5,000 Ref resamples with the predictions fixed. The methods are the same as in a, except that Gaussian process regression was not included. The log-scale local ridge regression includes the extreme prediction for Ref32. c, MAE of the calibrated random forests with the three input schemes and of the two local ridge regressions for Qiantang, shown as medians and interquartile ranges over the 200 draws (kPa). Paired error ratios and other calibration methods are given in Supplementary Fig. S5 and Supplementary Tables S14–S19.

3.3.2 Extreme predictions under extrapolation

Extreme predictions occurred when log-scale fits were extrapolated beyond the range of the local records (Supplementary Section S8). For AAC, the log-scale local ridge regression predicted 311,484.07 MPa in one draw for Ref32 (measured, 35.68 MPa), in which the total-water input of the test records lay outside the range of the local records. For P37 with the Base scheme and 5 local records, one log-linear calibration predicted 4.95×10^{14} kPa, and the source-model predictions for the

test records lay outside the range of those for the local records. For the 20% Qiantang task, the calibrated Gaussian process regression had a median MAE of 50,707.6 kPa, compared with 1,296.2 kPa without calibration (Supplementary Table S18).

4 Discussion

Under LOSO validation, the error of the random forest was 3.74 times that under the record random split. A similar gap was reported for AAC, where R^2 was 0.801 under a random split and had a median of −0.328 under leave-one-study-out validation across 85 studies [17]. Under a record random split, other records of the same source remain in the training set, and the accuracy reported by M'harzi Alaoui et al. [12] for a record-based split therefore refers to sources already represented in training. The overall MAE of the random forest rose from 133.9 kPa under the record random split to 200.4 kPa under the mixture-group split and 500.6 kPa under LOSO validation. Of the total increase of 366.7 kPa, 300.2 kPa (81.9%) appeared once entire sources were held out, indicating that the increase in new-material error arose mainly from differences between sources. Pacheco-Torgal and Iqbal [17] attributed the gap between random-split and leave-one-study-out metrics to the clustering of records within studies. The source-coverage comparison separated the effect of target-source records from differences in the number of training records and the test records. It gave a source-averaged MAE 2.57 times as high without target-source records in training as with them (1.77 times for AAC), which supports this attribution with training-set size and test records fixed. The multilayer perceptron had a lower median MAE than the random forest under the record random split, whereas the interval of their MAE difference under LOSO validation included zero (Supplementary Section S5). Error estimation and model comparison for new materials should therefore use LOSO validation.

Adding water content as an input lowered the overall error, yet the matched pairs showed predicted strength changes that were zero or differed in magnitude from the measured changes (Section 3.2.2). Because 97.75% of the total sum of squares of water content in the training data lay between sources, the water-content input carried differences between sources. Onyelowe et al. [9] concluded

590 from a sensitivity analysis of literature-compiled data that optimum water content had an effect on
591 UCS close to zero. When water content varies mainly between sources, such a sensitivity includes
592 differences between sources. Comparing pairs within the same batch of material controls for
593 material differences between sources. The strength of cement-stabilized clay decreases as the ratio
594 of clay water content to cement content increases [32,33]. For the water contents compared here, the
595 median measured ΔUCS was negative in all four batches, ranging from −1,770.0 to −8.7 kPa,
596 consistent with this relationship. Pairs with a predicted change of zero did not reproduce this
597 direction. A similar evaluation problem has been reported in molecular property prediction. There,
598 all tested models had higher errors on activity cliffs, structurally similar molecules whose activities
599 differ by more than tenfold, than on other molecules, and the authors recommended reporting cliff-
600 specific metrics alongside overall metrics [34]. Water-content effects should be tested with matched
601 pairs from the same material, and multi-source databases should include series of tests on the same
602 material at different water contents.

603 With 10 local records, the calibrated random forest had a lower median MAE than the three local
604 reference models for P37 and Li2023 and a higher median MAE than the two local ridge
605 regressions in the three Qiantang dosage tasks. Relative to the raw-scale local ridge regression, its
606 median paired MAE ratios were 0.93 for P37 and 0.80 for Li2023, and the interquartile range for
607 P37 included 1. For AAC, the calibrated random forest had a lower MAE than the local median and
608 a higher MAE than the raw-scale local ridge regression (Section 3.3.1). Of the 24 Qiantang records,
609 15 had strengths above the training maximum of 3,529 kPa, whereas all predictions of the
610 uncalibrated random forest stayed within the training range (Supplementary Section S6).
611 Calibration had to extrapolate the source-model predictions beyond the training strength range,
612 while the local ridge regressions were fitted directly to the Qiantang local records, consistent with
613 their lower errors for Qiantang. Gaussian process regression, which had the lowest LOSO MAE,
614 had a higher error than the raw-scale local ridge regression after calibration for P37, Li2023 and
615 Qiantang, and in the 20% Qiantang task its median MAE was 50,707.6 kPa, 39 times that without

calibration (Supplementary Table S18). The ranking of source models by direct-prediction error did not carry over to their relative performance after calibration. Devi et al. [21] used target data to fine-tune pretrained models and found an exception with 10 target records. We used 10 local records for log-linear calibration, and the calibrated source models had higher errors than the local reference models for some target materials, in line with this exception. Whether calibrating a source model suits a new material should be judged by comparing it with a local reference model built from the same local records, on the same test records.

This study is a retrospective comparison of 10 training sources and 4 target materials, and source-level inferences and interval estimates refer to these sources. P37 and Li2023 had been used in preliminary analyses. The strengths of 175 of the 399 training records (43.9%) were extracted from figures, and the reading tolerances and their effect on the errors are reported in Supplementary Section S1. The target-covered group in the source-coverage comparison was formed by replacing records of other sources, which changed the source composition of the training set, and the difference between the two groups therefore also includes the effect of this change. Because the calibrated source models and the local reference models differ in training data, model form and inputs, their error comparison reflects these combined differences. The Water scheme was selected from a variable-by-variable comparison on the same data. The comparisons of XGBoost, Gaussian process regression and the multilayer perceptron, the raw-scale local ridge regression and the leaf-node statistics were added after preliminary results had been obtained. The source-resampling intervals were calculated with the predictions fixed and reflect the variability from the composition of the existing sources. These results apply to the materials and retrospective prediction tasks examined here, and performance on other materials should be evaluated with new, independent sources.

5 Conclusions

Under LOSO validation, the overall MAE of the random forest was 500.6 kPa, 3.74 times the median overall MAE under the record random split, and the split error ratios of the six source

models ranged from 2.46 to 6.52. In the source-coverage comparison, with the number of training records and the test records fixed, the source-averaged MAE of the random forest was 2.57 times as high without target-source records in training as with them (1.77 times for AAC). Adding water content to the Base scheme lowered the source-averaged MAE from 679.2 to 362.1 kPa, 13.1% below that of the Full scheme. Matched water-content pairs from the same material showed that some models predicted zero strength change or a change that deviated in magnitude from the measured one. With 10 local records, the calibrated random forest had a lower median MAE than the three local reference models for P37 and Li2023 and a higher median MAE than the two local ridge regressions in the three Qiantang dosage tasks. For AAC, the calibrated random forest had a lower MAE than the local median and a higher MAE than the raw-scale local ridge regression. Gaussian process regression, which had the lowest LOSO error, had a higher error than the raw-scale local ridge regression after calibration for all three target materials. For a new material, prediction error should be assessed with LOSO validation, the effect of water content with matched water-content pairs from the same material, and the effect of calibration by comparing the calibrated source model with local reference models built from the same local records and evaluated on the same test records. Multi-source databases should retain the source of each record and include series of tests on the same material at different water contents.

Data availability

The stabilized fine-grained material data were compiled from the sources listed in Table 1 and Supplementary Table S2. The AAC data were obtained from the public database of Moreno Torres et al. [22,23] (Zenodo version 2, https://doi.org/10.5281/zenodo.7805018, CC BY 4.0). Principal summary data, calculation definitions and sensitivity results supporting the analysis are provided in the main text and Supplementary Information. The literature-extracted records, analysis inputs and saved results are available from the corresponding author upon reasonable request.

Code availability

Post-processing scripts, plotting code and associated run instructions are available from the corresponding author upon reasonable request.

References

[1] Qiu, C., Li, Y., Li, X., Xu, G. & Zhang, D. Effects of high curing pressure on the unconfined compressive strength of cement-stabilized bottom sediments with high water content. *Buildings* **15**, 2869 (2025).

[2] Jiang, P. et al. The time effect and micromechanism of the unconfined compressive strength of cement modified slurries. *Adv. Mater. Sci. Eng.* **2021**, 5597275 (2021).

[3] Carreira, C., Bollwerk, S. M. & Lønborg, C. A review on beneficial use of dredged marine sediment. *Anthropocene Coasts* **8**, 12 (2025).

[4] Song, C., Kong, J., Wang, L., Yang, Q. & Duan, Z. Quantifying carbon emissions from different urban slurry resource utilization pathways: a life cycle assessment approach. *Low-carbon Mater. Green Constr.* **4**, 5 (2026).

[5] Yoobanpot, N., Jamsawang, P., Simarat, P., Jongpradist, P. & Likitlersuang, S. Sustainable reuse of dredged sediments as pavement materials by cement and fly ash stabilization. *J. Soils Sediments* **20**, 3807–3823 (2020).

[6] Yoobanpot, N., Jamsawang, P., Poorahong, H., Jongpradist, P. & Likitlersuang, S. Multiscale laboratory investigation of the mechanical and microstructural properties of dredged sediments stabilized with cement and fly ash. *Eng. Geol.* **267**, 105491 (2020).

[7] Wang, D., Abriak, N. E. & Zentar, R. Strength and deformation properties of Dunkirk marine sediments solidified with cement, lime and fly ash. *Eng. Geol.* **166**, 90–99 (2013).

[8] Yin, C., Zhang, W., Jiang, X. & Huang, Z. Effects of initial water content on microstructure and mechanical properties of lean clay soil stabilized by compound calcium-based stabilizer. *Materials* **11**, 1933 (2018).

691 [9] Onyelowe, K. C. et al. Estimating the strength of soil stabilized with cement and lime at optimal
692 compaction using ensemble-based multiple machine learning. *Sci. Rep.* **14**, 15308 (2024).
693 [10] Kumar, A., Sinha, S. & Pradhananga, R. Assessment of unconfined compressive strength of
694 nano-doped fly ash-treated clayey soil using machine learning tools. *Sci. Rep.* **15**, 33847 (2025).
695 [11] Aggarwal, V. et al. Prediction of unconfined compressive strength of fly ash and waste paper
696 sludge stabilized clayey soil using explainable machine learning. *Sci. Rep.* **16**, 16376 (2026).
697 [12] M'harzi Alaoui, F. E. et al. Predicting the unconfined compressive strength of lime-stabilized
698 soils under wet-dry cycles using extreme gradient boosting optimized with metaheuristic
699 algorithms. *J. Eng. Appl. Sci.* **73**, 263 (2026).
700 [13] Roberts, D. R. et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or
701 phylogenetic structure. *Ecography* **40**, 913–929 (2017).
702 [14] Ploton, P. et al. Spatial validation reveals poor predictive performance of large-scale ecological
703 mapping models. *Nat. Commun.* **11**, 4540 (2020).
704 [15] Kapoor, S. & Narayanan, A. Leakage and the reproducibility crisis in machine-learning-based
705 science. *Patterns* **4**, 100804 (2023).
706 [16] Li, K., DeCost, B., Choudhary, K., Greenwood, M. & Hattrick-Simpers, J. A critical
707 examination of robustness and generalizability of machine learning prediction of materials
708 properties. *npj Comput. Mater.* **9**, 55 (2023).
709 [17] Pacheco-Torgal, F. & Iqbal, S. Machine learning for alkali-activated concrete: feature
710 attribution, strength–carbon relationships, and the limits of out-of-campaign generalisation. *Constr.*
711 *Mater.* **6**, 56 (2026).
712 [18] Kadhim, S. N., Kurtoğlu, A. E., Bakbak, D., Çevik, A. & Özçetin, A. İ. A stacking ensemble-
713 based framework for predicting the compressive strength of microwave-cured geopolymers.
714 *Materials* **19**, 2474 (2026).
715 [19] Liu, Y., Ong, D. E. L., Oh, E., Liu, Z. & Hughes, R. Sustainable cementitious blends for
716 strength enhancement of dredged mud in Queensland, Australia. *Geotech. Res.* **9**, 65–82 (2022).

[20] Vahedifard, F., Howard, I. L., Bazne, M., Smith, B. T. & Barksdale, M. A. *Sustainably Enhancing Intermodal Freight Operation of Ports Using Geotextile Tubes*. Report NCITEC 2013-05/CMRC-15-01 (Mississippi State University, 2015); https://rosap.ntl.bts.gov/view/dot/40015.

[21] Devi, R., Butler, K. T. & Sai Gautam, G. Optimal pre-train/fine-tune strategies for accurate material property predictions. *npj Comput. Mater.* **10**, 300 (2024).

[22] Moreno Torres, B., Völker, C. & Firdous, R. An alkali-activated concrete dataset for sustainable building materials (version 2). *Zenodo* https://doi.org/10.5281/zenodo.7805018 (2023).

[23] Moreno Torres, B., Völker, C. & Firdous, R. Concreting a sustainable future: a dataset of alkali-activated concrete and its properties. *Data Brief* **50**, 109525 (2023).

[24] Ramani, P. V. & Chinnaraj, P. K. Geopolymer concrete with ground granulated blast furnace slag and black rice husk ash. *Građevinar* **67**, 741–748 (2015).

[25] Venkatesan, R. P. & Pazhani, K. C. Strength and durability properties of geopolymer concrete made with ground granulated blast furnace slag and black rice husk ash. *KSCE J. Civ. Eng.* **20**, 2384–2391 (2016).

[26] Vinai, R., Rafeet, A., Soutsos, M. & Sha, W. The role of water content and paste proportion on physico-mechanical properties of alkali activated fly ash–ggbs concrete. *J. Sustain. Metall.* **2**, 51–61 (2016).

[27] Rafeet, A., Vinai, R., Soutsos, M. & Sha, W. Guidelines for mix proportioning of fly ash/GGBS based alkali activated concretes. *Constr. Build. Mater.* **147**, 130–142 (2017).

[28] Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).

[*29*] Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (ACM, 2016).

[30] Rasmussen, C. E. & Williams, C. K. I. *Gaussian Processes for Machine Learning* (MIT Press, Cambridge, MA, 2006).

[31] Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

[32] Miura, N., Horpibulsuk, S. & Nagaraj, T. S. Engineering behavior of cement stabilized clay at high water content. *Soils Found.* **41**, 33–45 (2001).

[33] Horpibulsuk, S., Miura, N. & Nagaraj, T. S. Assessment of strength development in cement-admixed high water content clays with Abrams' law as a basis. *Géotechnique* **53**, 439–444 (2003).

[34] van Tilborg, D., Alenicheva, A. & Grisoni, F. Exposing the limitations of molecular machine learning with activity cliffs. *J. Chem. Inf. Model.* **62**, 5938–5951 (2022).

[35] Gong, S. et al. Research on the impact of carbide slag content on the strength and microstructure of solidified sludge during composite excitation. *PLoS ONE* **19**, e0314809 (2024).

[36] Wang, S. et al. Investigation on water transformation and pore structure of cement-stabilized dredged sediment based on NMR technology. *Materials* **15**, 3178 (2022).

[37] Nontananandh, S. et al. Sustainable utilization of dredged sediments and water treatment sludges as construction materials through combined dewatering and cement stabilization techniques. *Recycling* **10**, 22 (2025).

[38] Ribeiro, I. S., Fagundes, D. d. F. & Nierwinski, H. P. Effects of ettringite formation on the stability of cement-treated sediments. *Resources* **14**, 73 (2025).

[39] Zeng, S., Ma, N. & Choi, C. E. Experimental investigation of alkali-activated GGBS stabilization of dredged marine deposits: mechanical performance and decarbonization. *Acta Geotech.* https://doi.org/10.1007/s11440-026-03006-5 (2026).

[40] Chompoorat, T., Thanawong, K. & Likitlersuang, S. Swell-shrink behaviour of cement with fly ash-stabilised lakebed sediment. *Bull. Eng. Geol. Environ.* **80**, 2617–2628 (2021).

[41] Li, J. S., Chen, X., Lang, L., He, X. X. & Xue, Q. Evaluation of natural and artificial fiber reinforcements on the mechanical properties of cement-stabilized dredged sediment. *Soils Found.* **63**, 101319 (2023).

[42] Zhang, L. et al. Unconfined compressive strength of cement-stabilized Qiantang River silty clay. *Materials* **17**, 1082 (2024).

[43] Zhu, X. et al. Effect of calcium carbide residue on strength development along with mechanisms of cement-stabilized dredged sludge. *Materials* **15**, 4453 (2022).

Acknowledgements

This work was supported by the Major Science and Technology Research and Development Project of Jiangxi Province (Grant No. 20232ACG01004), the National Science and Technology Major Project for Deep Earth Exploration (Grant No. 2025ZD1011200), and the Special Fund for Postgraduate Innovation of Jiangxi Province (Grant No. YC2025-B165).

Author contributions

Lu Shen: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft. **Tugen Feng:** Supervision, Resources, Funding acquisition, Writing – review & editing. **Ming Ma:** Validation.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.docx